Collecting reward to defend homeostasis: A homeostatic reinforcement learning theory

Mehdi Keramati and Boris Gutkin

bioRxiv first posted online May 14, 2014
Access the most recent version at doi: http://dx.doi.org/10.1101/005140

Collecting reward to defend homeostasis: A homeostatic reinforcement learning theory

Authors: Mehdi Keramati[1,2,*], Boris Gutkin[1,3,*]

Affiliations:

[1] Group for Neural Theory, INSERM U960, Departément des Etudes Cognitives, Ecole Normale Supérieure, 75005 Paris, France.

[2] Gatsby Computational Neuroscience Unit, University College London, London, UK.

[3] National Research University Higher School of Economics, Center for Cognition and Decision Making, Moscow, Russia.

*Correspondence to: Mehdi@gatsby.ucl.ac.uk or Boris.gutkin@ens.fr

Abstract: Efficient regulation of internal homeostasis and defending it against perturbations requires complex behavioral strategies. However, the computational principles mediating brain's homeostatic regulation of reward and associative learning remain undefined. Here we use a definition of primary rewards, as outcomes fulfilling physiological needs, to build a normative theory showing how learning motivated behavior is modulated by the internal state of the animal. The theory proves that seeking rewards is equivalent to the fundamental objective of physiological stability, defining the notion of physiological rationality of behavior. We further give a formal basis for temporal discounting of reward. It also explains how animals learn to act predictively to preclude prospective homeostatic challenges, and attributes a normative computational role to the modulation of midbrain dopaminergic activity by hypothalamic signals.

Introduction

Survival requires living organisms to maintain their physiological integrity within the environment. In other words, they must preserve homeostasis (e.g. body temperature, glucose level, etc.). Yet, how might an animal learn to structure its behavioral strategies to obtain the outcomes necessary to fulfill and even preclude homeostatic challenges? In this sense, efficient behavioral decisions depend on two brain circuits working in concert: the hypothalamic homeostatic regulation (HR) system, and the cortico-basal ganglia reinforcement learning (RL) mechanism. However, the computational mechanisms underlying this obvious coupling remain poorly understood, as the two systems have traditionally been studied separately.

On the one hand, classical negative feedback models of HR explain hypothalamic function in behavioral sensitivity to the "internal" state, by axiomatizing that animals minimize the deviation of some key physiological variables from their hypothetical setpoints (Marieb & Hoehn, 2012). To this end, a corrective response is triggered when a deviation from setpoint is sensed or anticipated (Sibly & McFarland, 1974; Sterling, 2012). A key lacuna in these models is how a simple corrective response (e.g. "go eat") in response to a homeostatic deficit should be translated into a complex behavioral strategy for interacting with the dynamic and uncertain external world.

On the other hand, the computational theory of RL successfully explains the role of the cortico-basal ganglia system in behavioral adaption to the "external" environment, by exploiting experienced environmental contingencies and reward history (Rangel, Camerer, & Montague, 2008; Sutton & Barto, 1998). Critically, this theory is built upon one major axiom, namely, that the objective of behavior is to maximize reward acquisition. Yet, this suite of theoretical models does not resolve how the brain constructs the reward itself, and how the variability of the internal state impacts overt behavior.

Neurobiologically, accumulating evidence indicates intricate intercommunication between the hypothalamus and the reward-learning circuitry (Palmiter, 2007; Yeo & Heisler, 2012). The integration of the two systems is also behaviorally manifest in the classical behavioral pattern of anticipatory responding, in which animals learn to act predictively to preclude prospective homeostatic challenges. Here, we suggest an answer to the question of what computations, at an algorithmic level, are being performed in this biological integration of the two systems. Behaviorally, the theory explains anticipatory responding, extinction burst, and the rise-fall pattern of the response rate, as three behavioral phenomena for which the interaction between the two systems is necessary; thus, neither classical RL nor classical HR theories can account for these phenomena.

Results

Reward as need fulfillment. The term “reward” (equivalently: reinforcer, utility) has been at the heart of behavioral psychology since its foundation. In behavioral terms, reward refers to a stimulus that strengthens a desired response. According to RL theory, given the rewarding value of each outcome, animals learn the value of alternative choices as they experience them and receive feedback (i.e., reward) from the environment. However, animal behavior is variable even under well-controlled external conditions, suggesting that outcomes depend on the internal state.

The interaction between drive (as an internal state) and reward has been the subject of considerable debate in the motivation literature in psychology. Neo-behaviorists like Hull (Hull, 1943), Spence (Spence, 1956) and Mowrer (Mowrer, 1960) proposed the “drive-reduction” theory of motivation to define the nature of reward. According to this theory, one primary mechanism underlying reward is the usefulness of the corresponding outcome in fulfilling the homeostatic needs of the organism (Cabanac, 1971). Inspired by this theory, we derive a formal definition of primary reward as the ability of an outcome to restore the internal equilibrium of the physiological state. In the following sections, we demonstrate that our formal elaboration of the drive-reduction theory alleviates the criticisms raised against it.

We first define "homeostatic space" as a multidimensional metric space in which each dimension represents one physiologically regulated variable (the horizontal plane in Figure 1). The physiological state of the animal at each time t can be represented as a point in this space, denoted by $H_t = (h_{1,t}, h_{2,t}, .., h_{N,t})$, where $h_{i,t}$ indicates the state of the i-th physiological variable. For example, $h_{i,t}$ can refer to the animal's glucose level, body temperature, plasma osmolality, etc. The homeostatic setpoint, as the ideal internal state, can be denoted as $H^* = (h_1^*, h_2^*, .., h_N^*)$. As a mapping from the physiological to the motivational state, we define the "drive" as the distance of the internal state from the setpoint (the three-dimensional surface in Figure 1):

$$D(H_t) = \sqrt[m]{\sum_{i=1}^{N} \left| h_i^* - h_{i,t} \right|^n} \qquad (1$$

Having defined drive, we can now provide a formal definition for primary reward based on drive reduction theory. Assume that as the result of an action, the animal receives an outcome o_t at time t. The impact of this outcome on different dimensions of the animal's internal state can be denoted by $K_t = (k_{1,t}, k_{2,t}, .., k_{N,t})$. For example, $k_{i,t}$ can be the quantity of glucose received as a result of outcome o_t. Hence, the outcome results in a transition of the physiological state from H_t to $H_{t+1} = H_t + K_t$ (see Figure 1) and thus, a transition of the drive state from $D(H_t)$ to

$D(H_{t+1}) = D(H_t + K_t)$. Accordingly, the rewarding value of this outcome can be defined as the consequent reduction of drive:

$$\begin{aligned} r(H_t, K_t) &= D(H_t) - D(H_{t+1}) \\ &= D(H_t) - D(H_t + K_t) \end{aligned} \tag{2}$$

Intuitively, the rewarding value of an outcome depends on the ability of its constituting elements to reduce the homeostatic distance from the setpoint. We propose that this definition of reward is used by the brain's reward learning machinery to structure behavior. Incorporating the physiological reward definition (Eq. 2) in a normative RL theory allows us to derive the major result of our theory, which is that the rationality of behavioral patterns is geared toward maintaining physiological stability.

Rationality of the theory. Our definition of reward reconciles the RL and HR theories in terms of their normative assumptions: reward acquisition and physiological stability are mathematically equivalent behavioral objectives (see Materials and methods for the proof). More precisely, given the proposed definition of reward and given that animals discount future rewards (Chung & Herrnstein, 1967), any behavioral policy, π, that maximizes the sum of discounted rewards (SDR) also minimizes the sum of discounted deviations (SDD) from the setpoint, and vice versa. This can be represented as follows:

$$if \;\; \gamma < 1: \qquad \underset{\pi}{\operatorname{argmin}}\, SDD(\pi) = \underset{\pi}{\operatorname{argmax}}\, SDR(\pi) \tag{3}$$

where γ is the discount factor. In this respect, reward acquisition sought by the RL system is an efficient means to guide an animal's behavior toward satisfying the basic objective of defending homeostasis. Thus, our theory suggests a physiological basis for the rationality of reward seeking.

In the domain of animal behavior, one fundamental question is why animals should discount rewards the further they are in the future. Our theory indicates that reward seeking without discounting ($\gamma = 1$) would not lead, and may even be detrimental, to physiological stability (see Materials and methods). More precisely, in the absence of discounting, the rewarding value of behavioral policies that change the internal state only depends on the initial and final internal states, regardless of its trajectory in the homeostatic space. Thus, when $\gamma = 1$, the values of any two behavioral policies with equal net shifts of the internal state are equal, even if one policy moves the internal state along the shortest path, whereas the other policy results in large deviations of the internal state from the setpoint and threatens survival. These results hold for any form of temporal discounting (e.g., exponential, hyperbolic). In this respect, for the first time to our knowledge, our theory provides a normative explanation for the necessity of temporal discounting of reward: to maintain internal stability, it is necessary to discount future rewards.

Anticipatory responding. A paradigmatic example of behaviors governed by the internal state is the anticipatory responses geared to preclude perturbations in regulated variables even before any physiological depletion (negative feedback) is detectable. Anticipatory eating and drinking that occur before any discernible homeostatic deviation (S C Woods & Seeley, 2002), anticipatory shivering in response to a cue that predicts the cold (Hjeresen, Reed, & Woods, 1986; Mansfield, Benedict, & Woods, 1983), and insulin secretion prior to meal initiation (S C Woods, 1991), are only a few examples of anticipatory responding.

One clear example of a conditioned homeostatic response is animals' progressive tolerance to ethanol-induced hypothermia. Experiments suggest that this tolerance is mediated by associative learning processes (Mansfield & Cunningham, 1980): temperature deviations caused by ethanol injections decreased over trials, when injections preceded (i.e., were predictable) by a distinctive cue. Interestingly, in extinction trials where the ethanol was omitted, the animal temperature exhibited a significant increase above normal on cue presentation (Figure 2 - figure supplement 1). This result indicates that this tolerance is a conditioned response

Conditioned tolerance to the alcohol-induced homeostatic challenge clearly demonstrates that whether increasing the body temperature at any given time is rewarding or punishing depends on the internal state dynamics. Thus, the lack of internal state modulation of the rewarding value of responses in the classical RL theory renders it unable to explain anticipatory responding. Here we demonstrate that the integration of HR and RL processes can account for this phenomenon.

The model results (Figure 2) show that if a tolerance response (preventive increase in body temperature) precedes an ethanol injection, it results in a smaller deviation of body temperature from the setpoint, compared to the absence of the tolerance response (compare panels e and f of Figure 2). Therefore, expressing the tolerance response is the optimal behavior in terms of minimizing homeostatic deviation and thus, maximizing reward. As a result, the model gradually learns to choose this optimal strategy upon observing the cue (Figure 2b,c). In other words, the model learns to predictively generate a tolerance response, triggered by the conditioned stimulus, to minimize temperature deviation (see Materials and methods and Table S1 for simulation details). Thus, the model explains that the optimal homeostatic maintenance policy is acquired by associative learning mechanisms.

Our theory implies that animals are capable of learning not only Pavlovian learning (e.g. shivering, or tolerance to ethanol), but also instrumental anticipatory responding (e.g., pressing a lever to receive warmth, in response to a cold-predicting cue. See Figure S3). This is in contrast to the theory of predictive homeostasis where anticipatory behaviors are only *reflexive* responses to the predicted internal state upon observing cues (Sterling, 2012; Stephen C Woods & Ramsay, 2007).

Extinction burst. Extinction is a procedure where the reinforcer maintaining a behavior is no longer provided upon performing that behavior. Although extinction eventually results in decreased response rate, a transient increase in the response rate is often observed in the early stages. This classical phenomenon, known as “extinction burst”, is demonstrated for a variety of reinforcers, in both animals and humans (Skinner, 1938). To date, no normative explanation for such seemingly irrational behavior exists.

We propose that extinction burst is a result of the interplay between the learning and homeostatic systems. According to our theory, at the beginning of the extinction phase, the animal still expects to receive the outcome upon pressing the lever (Figure 3). During this period, due to the absence of an outcome, the internal state drops below the setpoint. This slight homeostatic deviation is sufficient to induce a transient increase in the response rate (i.e., burst), because the animal expects to receive the outcome and offset the internal deviation. Later on, as the animal learns that pressing the lever does not lead to the outcome, the rate of responding decreases in spite of aggravating homeostatic deviation.

In this respect, our model predicts that not only extinction, but even a reduction in the magnitude of the outcome will result in a temporary burst of responding (Figure S4). This prediction is in contrast to the classical homeostatic regulation models (see Materials and methods).

Rescuing Hull? Critically, our model is inspired by the drive reduction theory of motivation, initially proposed by Clark Hull (Hull, 1943), which became the dominant theory of motivation in psychology during the 1940s and 1950s. However, major criticisms have been leveled against this theory more recently (Berridge, 2004; McFarland, 1969; Savage, 2000). Our formal theory alleviates these major faults. In the earlier sections, we demonstrated that our theory redresses the first major fault of the classical drive-reduction: its inability to explain anticipatory responding in which animals paradoxically voluntarily increase (rather than decrease) their drive deviation, even in the absence of any physiological deficit. We demonstrated how such apparently maladaptive responses are learned and result in optimal behavior, ensuring physiological stability.

Second, the drive reduction could not explain how secondary reinforcers (e.g., money, or a light that predicts food) gain motivational value, since they do not reduce the drive *per se*. Because our framework integrates an RL module with the HR reward computation, the drive reduction-induced reward of primary reinforcers can be readily transferred through the learning process to secondary reinforcers that predict them (i.e., Pavlovian conditioning) as well as to behavioral policies that lead to them (i.e., instrumental conditioning).

Similarly, our integrated theory is able to give a normative account for the motivational effect of the orosensory components associated with primary physiological outcomes. To do so, we posit that sensory properties of food and water provide the animal with an unbiased estimate, $\widehat{K}_t$, of their true post-ingestive effect, K_t, on the internal state. Such association between sensory and post-ingestive properties could have been developed through learning or evolutionary mechanisms (Breslin, 2013).

Based on this sensory approximation, the reinforcing value of food and water outcomes can be approximated as soon as they are sensed/consumed, without having to wait for the outcome to be digested and the drive to reduce. In other words, the only information required to compute the reward (and thus the reward prediction error) is the current physiological state (H_t) and the sensory-based approximation of the nutritional content of the outcome ($\widehat{K}_t$):

$$r\left(H_t, \widehat{K}_t\right) = D(H_t) - D\left(H_t + \widehat{K}_t\right) \qquad (4$$

This proposition is in accordance with the fact that dopamine neurons exhibit instantaneous, rather than delayed, burst activity in response to unexpected food rewards (Schneider, 1989). Clearly, the evolution of the internal state itself ($H_t \to H_t + K_t$) depends only on the actual (K_t) post-ingestive effects of the outcome.

This plausible sensory-estimate extension alleviates the other faults of drive reduction theory. Notably it explains the experimental fact that intravenous injection (and even intragastric intubation, in some cases) of food is not rewarding even though its drive reduction effect is equal to when it is ingested orally (Miller & Kessen, 1952) (*see also* (Ren et al., 2010)*)*: the post-ingestive effect of food is estimated by its sensory properties and thus, the reinforcing value of intravenously injected food that lacks sensory aspects is zero. In the same line, the theory explains that animals' motivation toward palatable foods, such as saccharine, that have no caloric content (and thus no need-reduction effect) is due to erroneous over-estimation of their drive-reduction capacity, misguided by their taste or smell.

A seminal series of experiments (McFarland, 1969) demonstrated that the reinforcing and satiating (i.e., need reduction) effects of drinking behavior, dissociable from one another, are governed by the orosensory and alimentary components of the water, respectively. Thirsty animals learned to peck at a key only when it resulted in oral, but bot intragastric (through a fistula) delivery of water (Figure 4 - figure supplement 1). Also, the response rate in the oral group initially increased but then gradually extinguished (rise-fall pattern; Figure 4 - figure supplement 1a). Simulating our theory (Figure 4) to account for this behavioral pattern (Figure 4a) clarifies that the ascending limb of the response curve represents a learning effect (Figure

4.c), whereas the descending limb is due to the internal state approaching the setpoint (i.e., satiation, as opposed "unlearning"; Figure 4e). Notably, classical RL models only explain the rise, and classical HR models only explain the fall. Furthermore, the simulation results show that the sensory component is crucial for approximating the drive-reduction reinforcing effect of water (Figure 4a,b). As above, the sensory-based approximation ($\widehat{K}_t$) of the alimentary effect of water in the oral and fistula cases is assumed to be equal to its actual effect (K_t) and zero, respectively (See Figure 4 - figure supplement 3 and 4, and Table S2 for simulation details).

Our theory also explains why satiation is independent of the sensory aspects of water and only depends on its post-ingestive effects. Experiments show that when different proportions of water were delivered via the two routes in different groups, satiation (i.e., suppression of pecking) only depended on the total amount of water ingested, regardless of the delivery route (McFarland, 1969). Our model accounts for these data (Figure 4 - figure supplement 2), since evolution of the internal state only depends on the actual water ingested. Therefore, reaching the setpoint (and thus suppression of pecking) does not depend on the oral/fistula proportion. We predict, however, that the proportion affects the speed of satiation: higher oral proportions result in higher rewarding values of pecking, which in turn results in faster responding and thus faster satiation (Figure 4 - figure supplement 2).

Behavioral plausibility of drive. The definition of the drive function (Eq. 1) in our model has two degrees of freedom: m and n are free parameters whose values determine the properties of the homeostatic space metric. Appropriate choice of m and n ($n > m > 2$) permits our theory to account for the following four key behavioral phenomena in a unified framework (see Materials and methods for mathematical derivations): (a) the potentiation of the reinforcing value of an appetitive outcome as its dose (K_t) increases (Figure S5); (b) the excitatory effect of the deprivation level on reinforcing value (i.e., food will be more rewarding when the animal is hungrier; Figure S6); (c) the inhibitory effect of irrelevant drives (Figure S7), which is consistent with a large body of behavioral experiments showing competition between different motivational systems so that as the deprivation level for one need increases, it inhibits the rewarding value of other outcomes that satisfy irrelevant motivational systems (e.g., calcium deprivation reduces the appetite for phosphorus and hunger inhibits sexual behavior, etc.; see (Dickinson & Balleine, 2002) for a review); (d) finally, the theory naturally captures the risk-aversive nature of behavior. The rewarding value in our model is a concave function of the corresponding outcome amplitude (the second derivative of the reward with respect to the dose of outcome is negative). It is well known that the concavity of the utility function is equivalent to risk aversion (Mas-Colell, Whinston, & Green, 1995). Indeed, the model shows (Figure S8) that when faced with two options with equal expected payoffs, the model learns to choose the more certain option as opposed to the risky one. In fact, as an evolutionary function of risk-avoidance behavior, our theory provides the intuition that when the expected physiological instability of two behavioral options are equal, organisms do not choose the risky option, because the severe, though unlikely, physiological instabilities that it can cause might be life-threatening.

Our unified explanation for the above behavioral patterns implies that they all arise from the functional form of the mapping from the physiological to the motivational state.

Neural substrates. Homeostatic regulation critically depends on sensing the internal state. In the case of energy regulation, for example, the arcuate nucleus of the hypothalamus integrates peripheral hormones including leptin, insulin, and ghrelin, whose circulating levels reflect the internal abundance of fat, abundance of carbohydrate, and hunger, respectively (Williams & Elmquist, 2012). In our model, the deprivation level has an excitatory effect on the rewarding value of outcomes and thus on the reward prediction error (RPE). Consistently, recent evidence indicates neuronal pathways through which energy state-monitoring peptides modulate the activity of midbrain dopamine neurons, which supposedly carry the RPE signal (Palmiter, 2007).

Namely, orexin neurons, which project from the lateral hypothalamus area to several brain regions, including the ventral tegmental area (VTA) (Sakurai et al., 1998) have been shown to have an excitatory effect on dopaminergic activity (Korotkova, Sergeeva, Eriksson, Haas, & Brown, 2003; Narita et al., 2006). Orexin neurons are responsive to peripheral metabolic signals as well as to the animal's deprivation level (Burdakov, Gerasimenko, & Verkhratsky, 2005), as they are innervated by orexigenic and anorexigenic neural populations in the arcuate nucleus where circulating peptides are sensed. Accordingly, orexin neurons are suggested to act as an interface between internal states and the reward learning circuit (Palmiter, 2007). In parallel with the orexinergic pathway, ghrelin, leptin and insulin receptors are also expressed on the VTA dopamine neurons, providing a further direct interface between the HR and RL systems. Consistently, whereas leptin and insulin inhibit dopamine activity, ghrelin has an excitatory effect (see (Palmiter, 2007) for a review).

The reinforcing value of food outcome (and thus the RPE signal) in our theory is not only modulated by the internal state, but also by the orosensory information that approximates the need-reduction effects. In this respect, endogenous opioids and μ-opioid receptors have long been implicated in the hedonic aspects of food, signaled by its orosensory properties. Systemic administration of opioid antagonists decreases subjective pleasantness rating and affective responses for palatable foods in humans (Yeomans & Wright, 1991) and rats (Doyle, Berridge, & Gosnell, 1993), respectively. Supposedly through modulating palatability, opioids also control food intake (Sanger & McCarthy, 1980) as well as instrumental food-seeking behavior (Cleary, Weldon, O'Hare, Billington, & Levine, 1996). For example, opioid antagonists decrease the breakpoint in progressive ratio schedules of reinforcement with food (Barbano, Le Saux, & Cador, 2009), whereas opioid agonists produce the opposite effect (Solinas & Goldberg, 2005). This reflects the influence of orosensory information on the reinforcing effect of food. Consistent with our model, these influences have mainly been attributed to the effect of opiates on increasing extracellular dopamine levels in the Nucleus Accumbens (NAc) (Devine, Leone, & Wise, 1993) through its action on μ-opioid receptors in the VTA and NAc (Noel & Wise, 1993; M. Zhang & Kelley, 1997).

Discussion

Theories of conditioning are founded on the argument that animals seek reward, while reward is defined as what animals seek. This apparently circular argument relies on the hypothetical and out-of-reach axiom of reward-maximization as the behavioral objective of animals. Physiological stability, however, is an observable fact. Here, we established a coherent mathematical theory where physiological stability is put as the basic axiom, and reward is defined in physiological terms. We demonstrated that reinforcement learning algorithms under such a definition of physiological reward lead to optimal policies that both maximize reward collection and minimize homeostatic needs. This argues for behavioral rationality of physiological integrity maintenance and further shows that temporal discounting of rewards is paramount for homeostatic maintenance. Furthermore, we demonstrated that such integration of the two systems provides normative explanations for several behavioral and neurobiological phenomena, including anticipatory responding, extinction burst, the rise-fall pattern of food-seeking response, and the modulation of midbrain dopaminergic activity by hypothalamic signals. We then extended the theory to incorporate orosensory information as an unbiased estimate of the post-ingestive effects of outcomes that is available instantaneously upon consumption. This extension of the theory allowed it to explain further behavioral patterns, rescue the classical drive-reduction theory, and shed light on the role of dopamine modulation by the opioid system.

From an evolutionary perspective, physiological stability and thus survival can themselves be seen as means of guaranteeing reproduction. These intermediate objectives can be even violated in specific conditions and be replaced with parental sacrifice. Still, homeostatic maintenance can explain a majority of motivated behaviors in animals. It is also noteworthy that our theory only applies to rewards that have a corresponding regulatory system. How to extend our theory to rewards without a corresponding homeostatic regulation system (e.g., social rewards, novelty-induced reward, etc.) remains a key challenge for the future.

Using internal state-independent reward/utility in the classical RL models and more generally in standard economic choice theory leads to the inevitable conclusion that maximizing utility is equivalent to maximizing the acquisition of appetitive outcomes. That is, more of a commodity equals more happiness. In opposition to this view, the rational decision in our model under certain circumstances is to choose small rather than large outcomes to avoid overshooting the setpoint (Figure S9). Furthermore, the theory explained several other key behavioral phenomena that stem from the interaction between the RL and HR systems: anticipatory responding, extinction burst, and the rise-fall pattern of responding.

Interestingly, a linear approximation of our proposed drive-reduction reward is equivalent to assuming that the rewarding value of outcomes is equal to the multiplication of the deprivation level and the magnitude of the outcome (see Materials and methods). In this respect, our model subsumes and provides a normative basis to the incentive salience theory (J. Zhang, Berridge, Tindell, Smith, & Aldridge, 2009) as well as other multiplicative forms of deprivation-modulated reward (e.g., decision field theory (Busemeyer, Townsend, & Stout, 2002), intrinsically motivated RL theory (Singh, Lewis, Barto, & Sorg, 2010), and MOTIVATOR theory (Dranias, Grossberg, & Bullock, 2008)), where reward increases as a linear function of deprivation level.

Whether the brain uses a nonlinear drive-reduction reward (as in Eq. 2) or a linear approximation of it (as in Eq. 4) can be examined experimentally. Assuming that an animal is in a slightly deprived state (Figure 5a), a linear model predicts that as the magnitude of the outcome increases, its rewarding value will increase linearly. A non-linear reward, however, predicts an inverted U-shaped utility function (Figure 5b). That is, the rewarding value of a large outcome can be negative, if it results in overshooting the setpoint.

In a nutshell, our theory incorporates a formal physiological definition of primary rewards into a novel homeostatically regulated reinforcement learning theory, allowing us to prove that economically rational behaviors ensure physiological integrity. At the same time our theory opens a unified approach behavioral pathologies (e.g., obesity, psychiatric diseases, drug addiction and compulsivity disorders) as allostatic dysregulation of the interactions between the motivational and learning brain axes.

References:

Barbano, M. F., Le Saux, M., & Cador, M. (2009). Involvement of dopamine and opioids in the motivation to eat: influence of palatability, homeostatic state, and behavioral paradigms. *Psychopharmacology*, *203*(3), 475–87.

Berridge, K. C. (2004). Motivation concepts in behavioral neuroscience. *Physiology & Behavior*, *81*(2), 179–209.

Breslin, P. A. S. (2013). An evolutionary perspective on food and human taste. *Current biology : CB*, *23*(9), R409–18. doi:10.1016/j.cub.2013.04.010

Burdakov, D., Gerasimenko, O., & Verkhratsky, A. (2005). Physiological changes in glucose differentially modulate the excitability of hypothalamic melanin-concentrating hormone and orexin neurons in situ. *The Journal of Neuroscience*, *25*(9), 2429–2433.

Busemeyer, J. R., Townsend, J. T., & Stout, J. C. (2002). Motivational underpinnings of utility in decision making: decision field theory analysis of deprivation and satiation. In S. Moore & M. Oaksford (Eds.), *Emotional cognition: from brain to behaviour* (pp. 197–218). Amsterdam: John Benjamins.

Cabanac, M. (1971). Physiological Role of Pleasure. *Science*, *173*(4002), 1103–1107.

Chung, S. H., & Herrnstein, R. J. (1967). Choice and delay of reinforcement. *Journal of the experimental analysis of behavior*, *10*(1), 67–74. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1338319&tool=pmcentrez&rendertype=abstract

Cleary, J., Weldon, D. T., O'Hare, E., Billington, C., & Levine, A. S. (1996). Naloxone effects on sucrose-motivated behavior. *Psychopharmacology*, *126*(2), 110–4.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Devine, D. P., Leone, P., & Wise, R. A. (1993). Mesolimbic dopamine neurotransmission is increased by administration of mu-opioid receptor antagonists. *European journal of pharmacology*, *243*(1), 55–64.

Dickinson, A., & Balleine, B. W. (2002). The role of learning in motivation. In C. R. Gallistel (Ed.), *Volume 3 of Steven's Handbook of Experimental Psychology: Learning, Motivation, and Emotion* (3rd ed., pp. 497–533). New York: Wiley.

Doyle, T. G., Berridge, K. C., & Gosnell, B. A. (1993). Morphine enhances hedonic taste palatability in rats. *Pharmacology, biochemistry, and behavior*, *46*(3), 745–9.

Dranias, M. R., Grossberg, S., & Bullock, D. (2008). Dopaminergic and non-dopaminergic value systems in conditioning and outcome-specific revaluation. *Brain research*, *1238*, 239–87.

Hjeresen, D. L., Reed, D. R., & Woods, S. C. (1986). Tolerance to hypothermia induced by ethanol depends on specific drug effects. *Psychopharmacology*, *89*(1), 45–51.

Hull, C. L. (1943). *Principles of behavior: an introduction to behavior theory*. New York: Appleton-Century-Crofts.

Korotkova, T. M., Sergeeva, O. A., Eriksson, K. S., Haas, H. L., & Brown, R. E. (2003). Excitation of ventral tegmental area dopaminergic and nondopaminergic neurons by orexins/hypocretins. *The Journal of Neuroscience*, *23*(1), 7–11.

Mansfield, J. G., Benedict, R. S., & Woods, S. C. (1983). Response specificity of behaviorally augmented tolerance to ethanol supports a learning interpretation. *Psychopharmacology*, *79*(2-3), 94–98.

Mansfield, J. G., & Cunningham, C. L. (1980). Conditioning and extinction of tolerance to the hypothermic effect of ethanol in rats. *Journal of Comparative and Physiological Psychology*, *94*(5), 962–969.

Marieb, E. N., & Hoehn, K. (2012). *Human Anatomy & Physiology* (9th ed., p. 1264). Benjamin Cummings.

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic Theory*. Cambridge: Cambridge Univ. Press.

McFarland, D. (1969). Separation of satiating and rewarding consequences of drinking. *Physiology & Behavior*, *4*(6), 987–989. Retrieved from http://www.sciencedirect.com/science/article/pii/0031938469900547

Miller, N. E., & Kessen, M. L. (1952). Reward effects of food via stomach fistula compared with those of food via mouth. *Journal of Comparative and Physiological Psychology*, *45*(6), 555–564.

Mowrer, O. H. (1960). *Learning theory and behavior*. New York: Wiley.

Narita, M., Nagumo, Y., Hashimoto, S., Narita, M., Khotib, J., Miyatake, M., Sakurai, T., et al. (2006). Direct involvement of orexinergic systems in the activation of the mesolimbic dopamine pathway and related behaviors induced by morphine. *The Journal of neuroscience*, *26*(2), 398–405.

Noel, M. B., & Wise, R. A. (1993). Ventral tegmental injections of morphine but not U-50,488H enhance feeding in food-deprived rats. *Brain research*, *632*(1-2), 68–73.

Palmiter, R. D. (2007). Is dopamine a physiologically relevant mediator of feeding behavior? *Trends in neurosciences*, *30*(8), 375–81.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature reviews. Neuroscience*, *9*(7), 545–56.

Ren, X., Ferreira, J. G., Zhou, L., Shammah-Lagnado, S. J., Yeckel, C. W., & De Araujo, I. E. (2010). Nutrient selection in the absence of taste receptor signaling. *The Journal of Neuroscience*, *30*(23), 8012–23.

Sakurai, T., Amemiya, A., Ishii, M., Matsuzaki, I., Chemelli, R. M., Tanaka, H., Williams, S. C., et al. (1998). Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell*, *92*(5), 573–585.

Sanger, D. J., & McCarthy, P. S. (1980). Differential effects of morphine on food and water intake in food deprived and freely-feeding rats. *Psychopharmacology*, *72*(1), 103–6.

Savage, T. (2000). Artificial motives: A review of motivation in artificial creatures. *Connection Science*, *12*(3-4), 211–277. doi:10.1080/095400900750060131

Schneider, L. H. (1989). Orosensory self-stimulation by sucrose involves brain dopaminergic mechanisms. *Annals of the New York Academy of Sciences*, *575*, 307–19.

Sibly, R. M., & McFarland, D. J. (1974). *State Space Approach to Motivation, Motivational Control System Analysis*. Academic Press.

Singh, S., Lewis, R. L., Barto, A. G., & Sorg, J. (2010). Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective. *IEEE Transactions on Autonomous Mental Development*, *2*(2), 70–82.

Skinner, B. F. (1938). *The Behavior of Organisms*. New York: Appleton-Century-Crofts.

Solinas, M., & Goldberg, S. R. (2005). Motivational effects of cannabinoids and opioids on food reinforcement depend on simultaneous activation of cannabinoid and opioid systems. *Neuropsychopharmacology*, *30*(11), 2035–45.

Spence, K. W. (1956). *Behavior theory and conditioning*. Westport: Greenwood Press.

Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology & behavior*, *106*(1), 5–15.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge: MIT Press.

Williams, K. W., & Elmquist, J. K. (2012). From neuroanatomy to behavior: central integration of peripheral signals regulating feeding behavior. *Nature neuroscience*, *15*(10), 1350–5.

Woods, S C. (1991). The eating paradox: how we tolerate food. *Psychological Review*, *98*(4), 488–505.

Woods, S C, & Seeley, R. J. (2002). Hunger and energy homeostasis. In C. R. Gallistel (Ed.), *Volume 3 of Steven's Handbook of Experimental Psychology: Learning, Motivation, and Emotion* (3rd ed., pp. 633–68). New York: Wiley.

Woods, Stephen C, & Ramsay, D. S. (2007). Homeostasis: beyond Curt Richter. *Appetite*, *49*(2), 388–398.

Yeo, G. S. H., & Heisler, L. K. (2012). Unraveling the brain regulation of appetite: lessons from genetics. *Nature neuroscience*, *15*(10), 1343–9.

Yeomans, M. R., & Wright, P. (1991). Lower pleasantness of palatable foods in nalmefene-treated human volunteers. *Appetite*, *16*(3), 249–59.

Zhang, J., Berridge, K. C., Tindell, A. J., Smith, K. S., & Aldridge, J. W. (2009). A Neural Computational Model of Incentive Salience. *PLoS computational biology*, *5*(7).

Zhang, M., & Kelley, A. E. (1997). Opiate agonists microinjected into the nucleus accumbens enhance sucrose drinking in rats. *Psychopharmacology*, *132*(4), 350–60.

Acknowledgments: We thank Peter Dayan, Amir Dezfouli, Serge Ahmed, and Mathias Pessiglione for critical discussions and Oliver Hulme for commenting on the manuscript. This study was supported by fundings from Frontiers du Vivant, the French MESR, CNRS, INSERM, ANR, ENP and NERF. Support from the Basic Research Program of the Russian National Research University Higher School of Economics is gratefully acknowledged by BG.

Figures:

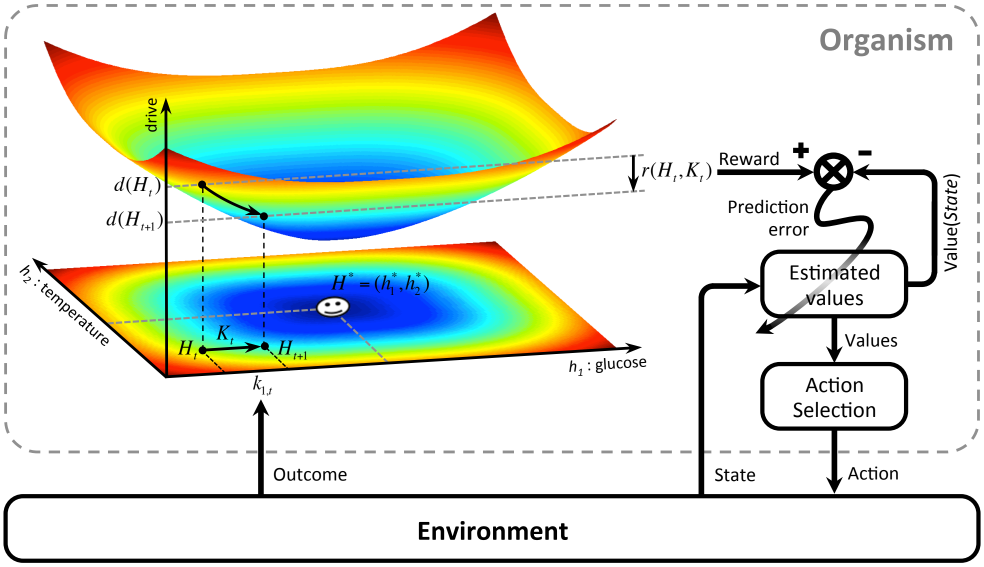

Figure 1. Schematics of the model in an exemplary two-dimensional homeostatic space. Upon performing an action, the animal receives an outcome K_t from the environment. The rewarding value of this outcome depends on its ability to make the internal state, H_t, closer to the homeostatic setpoint, H^*, and thus reduce the drive level (the vertical axis). This experienced reward, denoted by $r(H_t, K_t)$, is then learned by a RL algorithm. Here a model-free RL algorithm (Daw, Niv, & Dayan, 2005) is shown in which a reward prediction error signal is computed by comparing the realized reward and the expected rewarding value of the performed response. This signal is then used to update the subjective value attributed to the corresponding response. Subjective values of alternative choices bias the action selection process.

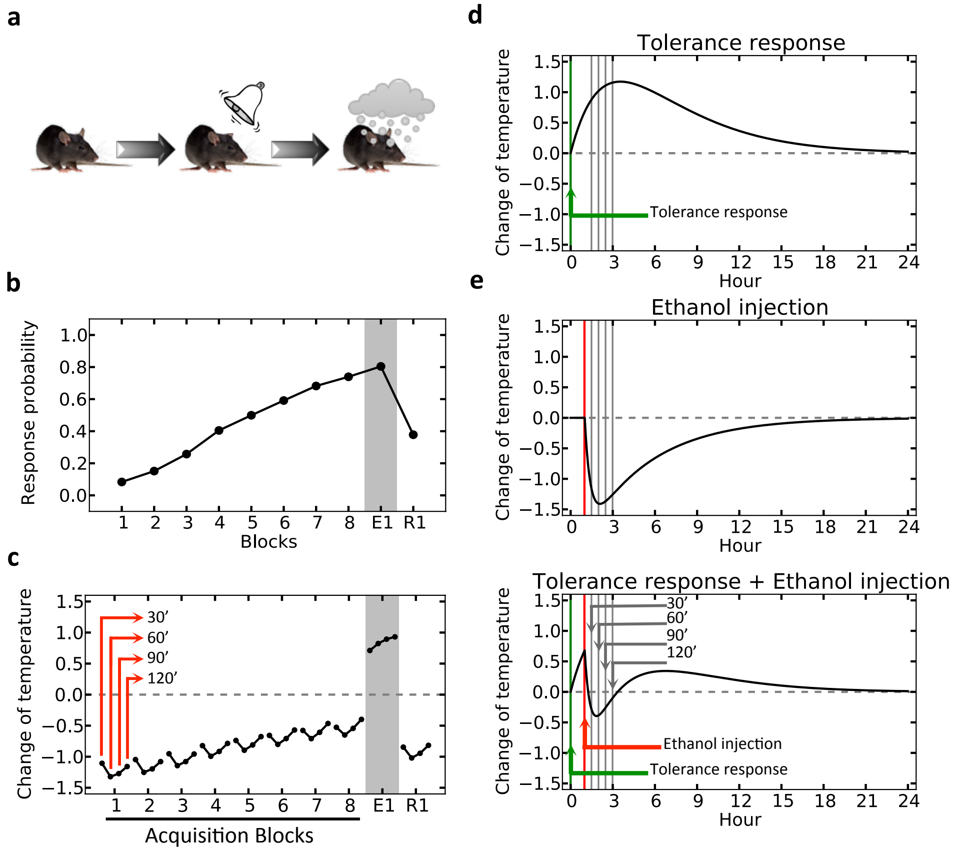

Figure 2. Anticipatory responding explained by the model. (a) In each block (day), the simulated agent receives an ethanol injection after the presentation of the stimulus. Plots (d) and (e) show the assumed dynamics of body temperature upon the initiation of the tolerance response and ethanol administration, respectively. Plot (f) shows the combined effect. Plot (b), which is averaged over 500 simulated agents, illustrates that the model gradually learns to choose the tolerance response after observing the stimulus. If the stimulus is not followed by ethanol injection, as in the first day of extinction (E1), it still triggers the tolerance response. However, the tolerance response is weakened after several extinction sessions, resulting in low response probability in the first day of re-acquisition (R1), where presentation of the cue is again followed by ethanol injection. (c) As in the experiment, the change in the body temperature is measured in

every block, 30, 60, 90, and 120 min after ethanol administration. Plot (c) replicates experimental results (Figure S3b).

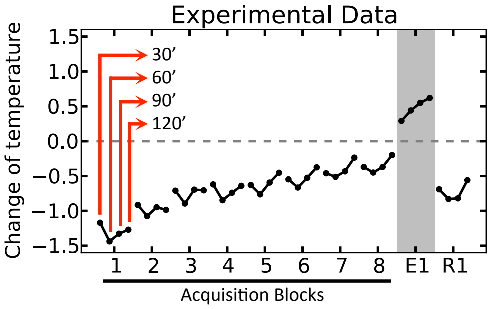

Figure 2 - figure supplement 1. Experimental results (adapted from *(18)*) on the acquisition and extinction of conditioned tolerance response to ethanol. In each block (day) of the experiment, the animal receives ethanol injection after the presentation of the stimulus. The change in the body temperature is measured 30, 60, 90, and 120 minutes after ethanol administration. Initially, the hypothermic effect of ethanol decreases the body temperature of animals. After several training days, however, animals learn to activate a tolerance response upon observing the stimulus, resulting in smaller deviations from the temperature setpoint. If the stimulus is not followed by ethanol injection, as in the first day of extinction (E1), the activation of the conditioned tolerance response results in an increase in body temperature. The tolerance response gets weakened after several extinction sessions, resulting in increased deviation from the setpoint in the first day of re-acquisition (R1), where presentation of the cue is again followed by ethanol injection.

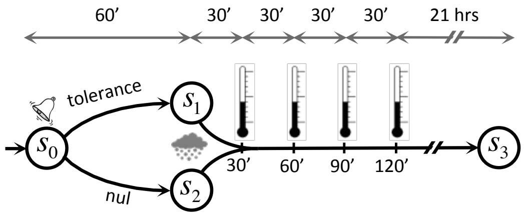

Figure 2 - figure supplement 2. The model is simulated in an artificial task (Markov Decision Process) in which, upon observing the stimulus, the agent can choose between triggering the tolerance response and doing nothing.

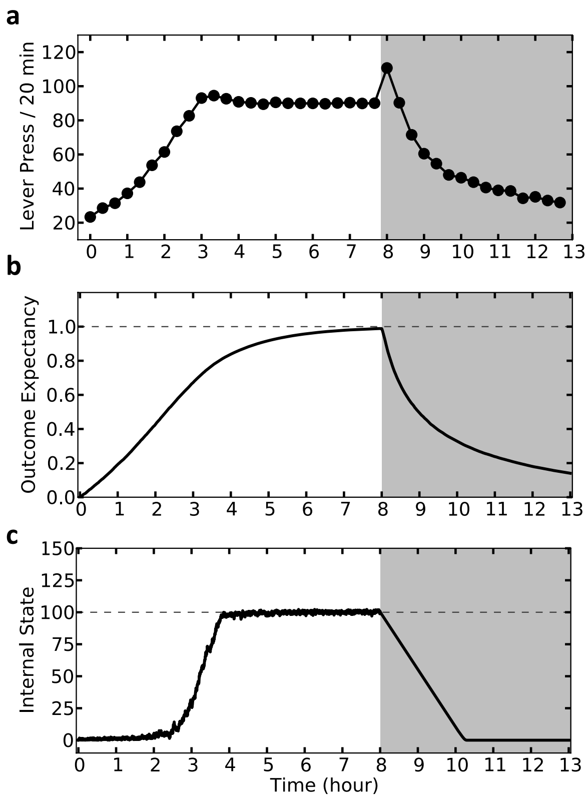

Figure 3. Simulation results demonstrating extinction burst. In every time step (every 4 s), the simulated agent can choose between pressing the lever (LP) or doing nothing (Figure S4). Pressing the lever results in an outcome with magnitude $k = 5$. Furthermore, in every time step, the internal state declines marginally, supposedly due to normal metabolism. The initial increase in response rate (a) is due to learning that pressing the lever results in an outcome (b). After the internal state reaches the setpoint, a stable response rate is maintained to preserve homeostasis (c). After 8 h, the outcome is removed, resulting in extinction burst followed by gradual suppression of responding.

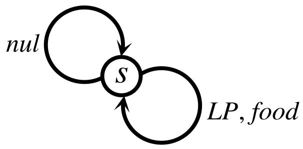

Figure 3 - figure supplement 1. The Markov Decision Process used for simulating the extinction burst. At each time point, the agent can choose between doing nothing (*nul*) or pressing the lever (LP) which results in food delivery.

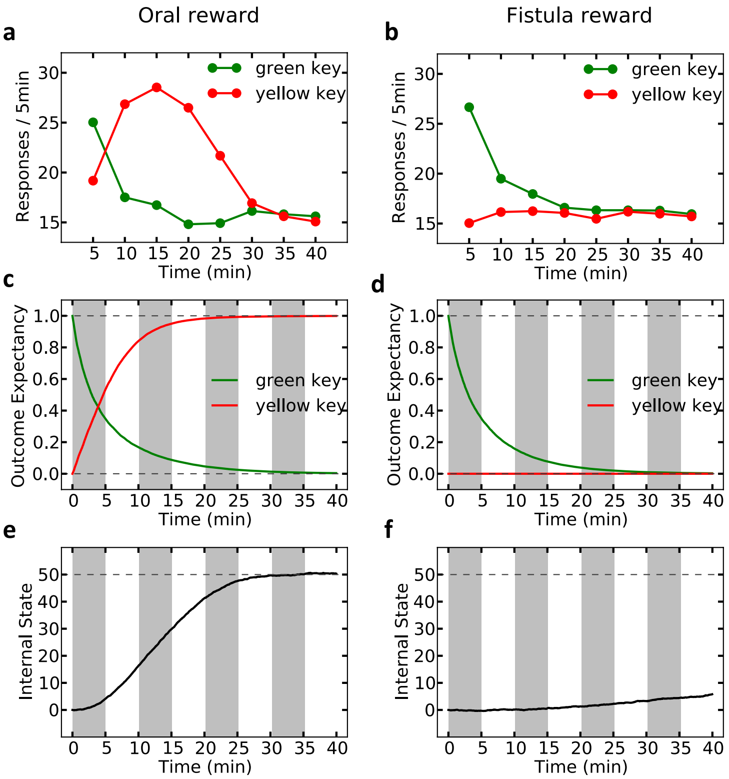

Figure 4. Simulation results replicating the data from (McFarland, 1969) on learning the reinforcing effect of oral vs. intragastric delivery of water. In a pre-training phase, simulated agents were trained in a task where responding on the green key resulted in the oral delivery of water. In the test phase, the green key had no consequence, whereas pecking at a novel yellow key resulted in oral delivery of water in one group (a) and intragastric injection of the same amount of water through a fistula in a second group (b). All agents started this phase in a thirsty state (initial internal state = 0; setpoint = 50). In the oral group, responding transferred rapidly from the green to the yellow key and was then suppressed (a) as the internal state approached the setpoint (e). This transfer is due to gradually updating the subjective probability of receiving water outcome upon responding on either key (c). In the fistula group, as the water is not sensed, the outcome expectation converges to zero for both keys (d) and thus, responding is extinguished (b). As a result, the internal state changes only slightly (f).

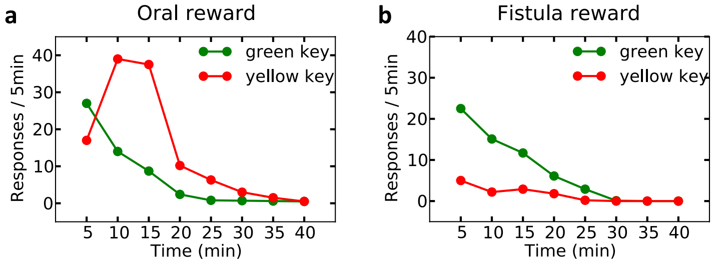

Figure 4 - figure supplement 1. Experimental results (adapted from *(23)*) on learning the reinforcing effect of oral vs. intragastric delivery of water. Thirsty animals were initially trained to peck at a green key to receive water orally. In the next phase, pecking at the green key had no consequence, while pecking at a novel yellow key resulted in oral delivery of water in one group (a), and intragastric injection of the same amount of water through a fistula in a second group (b). In the first group, responding was rapidly transferred from the green to the yellow key, and then suppressed. In the fistula group, the yellow key was not reinforced.

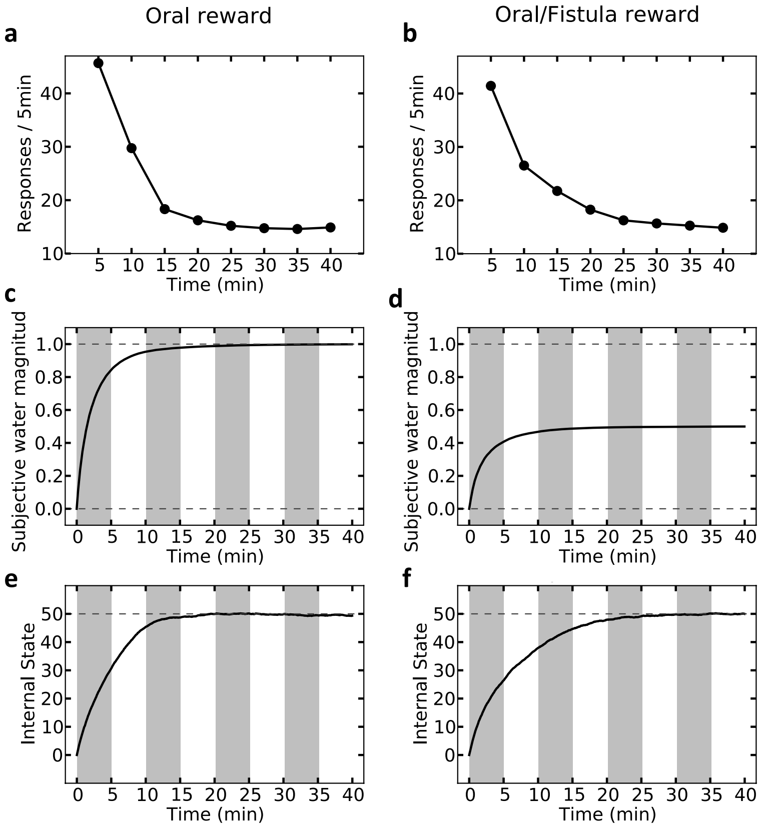

Figure 4 - figure supplement 2. Simulation results of the satiation test. Left column shows results for the case where water was received only orally. Rate of responding drops rapidly (a) as the internal state approaches the setpoint (e). Also, the agent learns rapidly that upon every key pecking, it receives 1.0 unit of water (c). On the right column, upon every key-peck, 0.5 unit of water is received orally, and 0.5 unit is received via the fistula. As only oral delivery is sensed by the agent, the subjective outcome magnitude converges to 0.5 (d). As a result, the reinforcing value of key-pecking is less than the oral case and thus, the rate of responding is lower (b). This in turn results in slower convergence of the internal state to the setpoint (f).

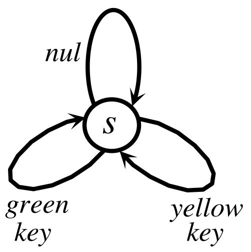

Figure 4 - figure supplement 3. The Markov Decision Process used for simulating the reinforcing vs. satiation effects of water. At each time point, the agent can choose between doing nothing (*nul*) or pecking at either the green or the yellow key.

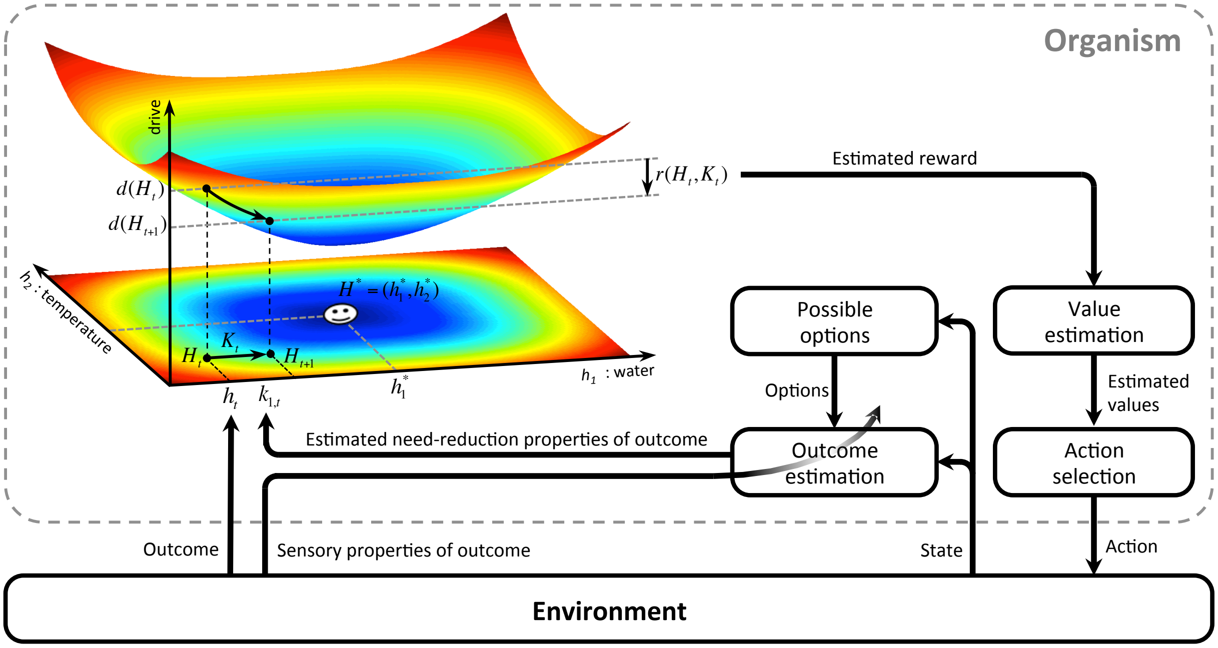

Figure 4 - figure supplement 4. A model-based homeostatic RL system. Upon performing an action in a certain state, the agent receives an outcome, K_t, which results in the internal state to shift from H_t to $H_t + K_t$. At the same time, sensory properties of the outcome are sensed by the agent. Based on this information, the agent updates the state-action-outcome associations. In fact, the agent learns to predict the sensory properties, $\widehat{K}_t$, of the outcome that is expected to be received upon performing a certain action. Having learned these associations, the agent can estimate the rewarding value of different options. That is, when the agent is in a certain state, it predicts the outcome $\widehat{K}_t$, expected to result from each behavioral policy. Based on $\widehat{K}_t$ and the internal state H_t, the agent can approximate the drive-reduction reward.

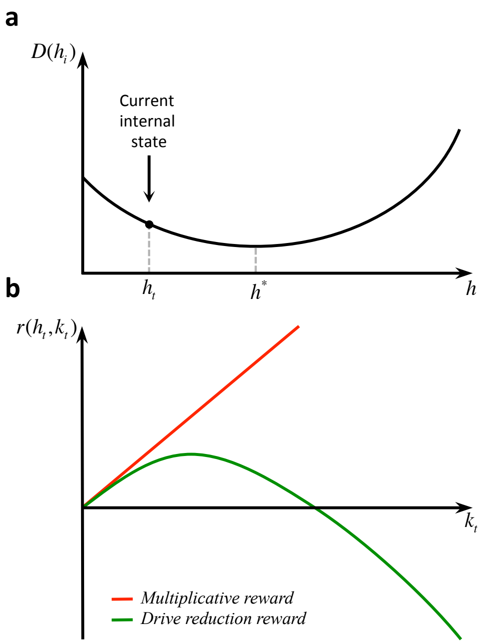

Figure 5. Differential behavioral predictions of multiplicative and drive-reduction forms of reward. Assuming that the internal state is at h_t (a), outcomes larger than $h^* - h_t$ result in overshooting the setpoint and thus a declining trend of rewarding value (b).